# HouseTour: A Virtual Real Estate A(I)gent: Appendix


## Abstract

*In the supplementary material, we provide:*

1. *Details on the HouseTour dataset (Sec. 1)*
2. *More qualitative results (Sec. 2)*
3. *User Evaluation of Text Generation (Sec. 3)*
4. *Implementation details (Sec. 4)*
5. *Out-of-Distribution Scenes (Sec. 5)*
6. *Details on the Metric Scale Evaluation (Sec. 6)*
7. *Details on the Bradley-Terry evaluation (Sec. 7)*
8. *Recall Plots for the trajectory generation (Sec. 8)*
9. *Preliminary information on diffusion (Sec. 9)*


## 1. HouseTour Dataset

### 1.1. Creation Details

At the start of our reconstruction pipeline, we select a subset of video frames for use in the process. Since our videos range from a few minutes to 15 minutes in length, using all frames would be computationally impractical. Our objective is to choose a minimal yet effective set of keyframes that maintain significant overlap to ensure accurate reconstruction. To accomplish this, we use an algorithmic approach that evaluates factors such as optical flow and keypoint matches between consecutive views.

We then trim the beginning and end of the selected keyframe sequence to remove frames that induce spatial jumps—such as exterior drone shots, which could largely affect the 3D reconstruction. To accomplish this, we use an off-the-shelf vision-language model, BLIP2 [3], to classify the keyframes at the sequence boundaries as exterior shots.

For the 3D scene reconstruction, we employ the COLMAP [5] structure-from-motion approach. We generate image pairs from the keyframes leveraging their inherent sequential order and augment these pairs with additional ones identified through traditional image retrieval techniques [6] to simulate loop closure during the reconstruction process. We perform dense 2D-to-2D matching between paired frames using Mast3r [2] and subsequently map them with COLMAP after geometrically verifying the pixel correspondences.

Lastly, if a video contains speech, we use Whisper [4] to extract the transcriptions along with their timestamps. If there is no speech, we obtain video descriptions as they stylistically align with the transcribed scene descriptions. To protect privacy, we employ GPT-4o to automatically filter out sensitive details such as addresses, personal names, and phone numbers from the video transcriptions. Additionally, we manually edit sections that mention neighborhood information or amenities, since such details are not visually represented in the videos.

### 1.2. Dataset Statistics

**3D Reconstruction and Scene-Level Descriptions.** The 3D reconstruction process for a single scene can take up to 40 hours, depending on the number of keyframes extracted from the videos. Reconstructing over 1,600 scenes may require 3 to 4 months of CPU time. Additionally, the keyframe extraction and matching processes are GPU-intensive. To manage this, we employ a high-performance computing cluster for dataset acquisition. Collecting the dataset typically takes about 7 days when using a job array with 20 parallel jobs. Each job requires 48GB of CPU memory and a 32GB NVIDIA Tesla V100 GPU. To handle extremely long runtimes, we limit each job to a maximum duration of 40 hours.

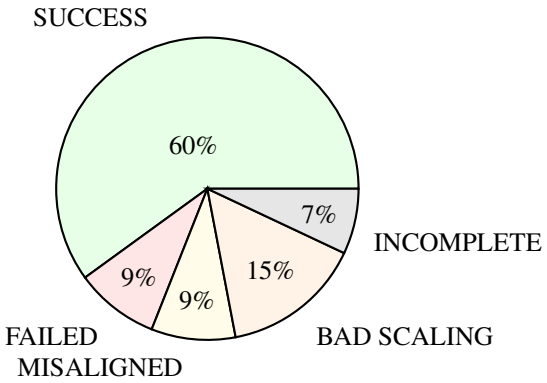



Table 1. **Breakdown of the reconstruction outcomes.**

Table 1 provides a detailed analysis of our pipeline's reconstruction performance. Scenes marked as *SUCCESS* have been reconstructed successfully. *INCOMPLETE* refers to reconstructions that failed to register the entire scene due to tracking loss, often caused by textureless areas or the lack of complete covisibility graph data. *MISALIGNED* scenes have errors in reconstruction leading to incorrect rotations of some scene portions. Scenes labeled with *BAD SCALING* have errors resulting in discrepancies in scale across parts of the output model. We classify scenes as *FAILED* if they exhibit multiple of the aforementioned issues or have significant errors in the final model.

Each reconstructed scene includes a dense point cloud with more than one million vertices, 2D-to-3D correspondences, and outputs from COLMAP [5], including images, camera data, and 3D points in binary files. Additionally, the selected keyframes and their timestamps are listed. For scenes with descriptions, we provide either a CSV file containing text and timestamp information or a plain text file with the description.

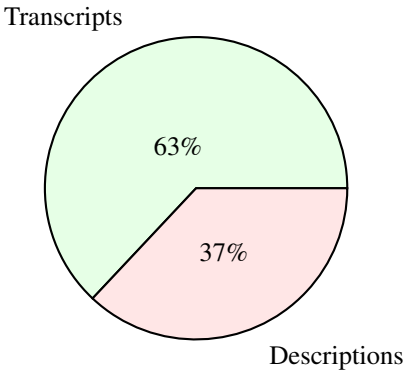



Table 2. **Distribution of summary types in the HouseTour dataset.** Transcripts include timestamped information.

Furthermore, we retrieved real estate descriptions for 1,298 out of 1,639 scenes. Of these, 813 descriptions were transcribed using the Whisper model, while the remaining ones were sourced from descriptions. We include all scenes in the dataset, even if they do not have 3D reconstructions or summaries. We utilize the scenes without 3D reconstruction as additional data to the VLM and those without summaries as additional data for the camera trajectory generation.

**Contextual Analysis of the Dataset Descriptions.** The videos primarily feature tours of detached houses rather than flats or apartments, with most properties located in a city within a developed, financially advanced country and its surrounding suburbs (we refrain from disclosing the location for privacy reasons). The showcased real estate is predominantly high-end, often including luxurious features. As shown in Table 3, these homes typically have three to five, with four being the most common. A substantial portion are multi-storey, usually with two or three floors, and many include outdoor spaces such as front or backyards. The interior design is largely modern or contemporary. It is important to note that our dataset is biased toward upscale properties in a specific region and does not capture the full diversity of global architectural styles.

**Linguistic Analysis of the Dataset Descriptions.** Table 4 presents an analysis of phrases based on constituency within the extracted descriptions. The constituency-based analysis reveals that the most common adjectives in the descriptions fall into categories such as scalar (e.g., "large", "double", and "ample"), directional (e.g., "main", "upper", and "lower"), and conceptual (e.g., "natural", "open", and "spacious"), all of which describe aspects of interior spaces. Additionally, there are adjectives related to material information, like "stainless" and "ceramic".

When analyzing the most frequent adverbs in the video descriptions, the word "fully" stands out as the most frequent adverb, occurring approximately 80 times, which may suggest these videos often showcase homes that are fully equipped or fully furnished. Following "fully," we see a significant mention of "beautifully," which indicates a focus on aesthetic appeal, highlighting beautifully designed spaces. Other adverbs like "away", "incredibly", and "graciously", suggest descriptions of location, extraordinary features, or hospitality aspects. Adverbs such as "professionally", "highly", and "conveniently" point towards emphasizing quality and ease of living. The occurrence of these specific adverbs suggests that house tour videos prioritize aspects such as completeness, beauty, functionality, and unique features to attract potential buyers or viewers.

Lastly, the verbs within the descriptions give further linguistic cues on descriptions as a whole. The verb "features" appears most frequently, over 400 times, indicating a focus on highlighting key aspects or amenities of properties. Verbs like "finished", "built", and "found" suggest emphasis on quality, construction, and location. Words such as "offered", "opens", and "overlooking" reflect the dynamic aspects and views these properties provide. Other verbs like "leads", "includes", and "showcases" emphasize navigation, inclusivity, and presentation within the space. The use of "situated", "covered", and "updated" implies a focus on positioning, protection, and modern enhancements. Overall, these verbs underline the importance of showcasing distinctive features and conveying a sense of completeness and modernization in house tours.

**Named-Entity Based Analysis of the Dataset Descriptions.** When analyzing the first bar plot in Table 5 the *rooms and areas in a house* entities in video descriptions, the "kitchen" emerges as the most highlighted space, with nearly 250 mentions, underscoring its role as a central and significant feature in homes. The "master bedroom" follows closely, reflecting its importance in personal comfort and privacy. Other commonly referenced areas include the "family room," "front foyer," and "living room," which are key spaces for gathering and welcoming guests. The plot also shows notable mentions of functional areas like the "laundry room" and "dining room," indicating their relevance in daily living. The presence of terms like "lower level" and "breakfast area" suggests an emphasis on specific sections or niches that add value to the property's layout. Overall, the focus on both communal and private spaces highlights a balanced presentation of essential living areas and unique features in house tours.

The second plot shows the frequency of *objects in a room*. "Stainless steel appliances" are the most prominent, appearing over 60 times, emphasizing the modern and desirable features of kitchens. Following closely are "granite countertops" and "gas fireplaces," indicating a

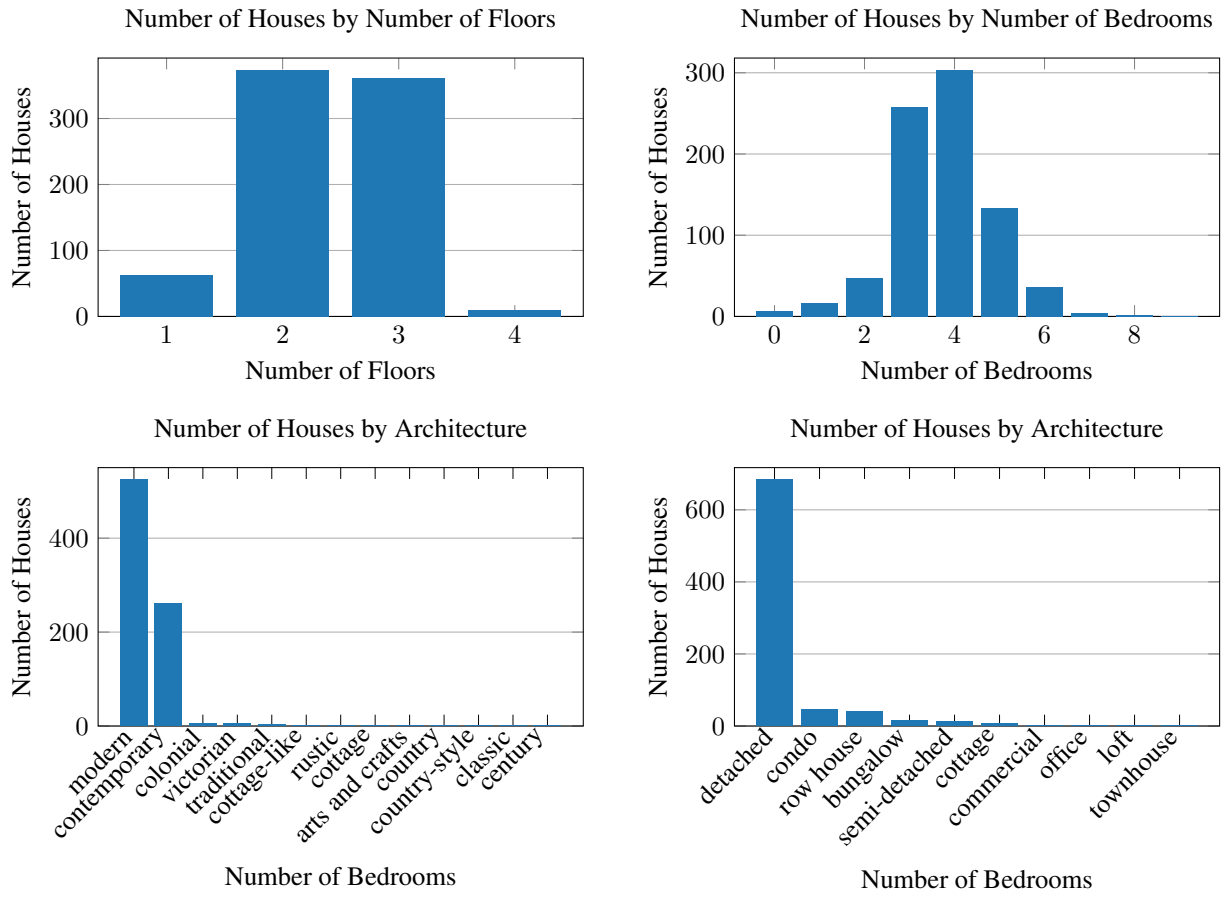



Table 3. **Contextual dataset statistics.** As shown in the data, most of the properties are modern, detached, with 2-3 floors, and 3-4 bedrooms.

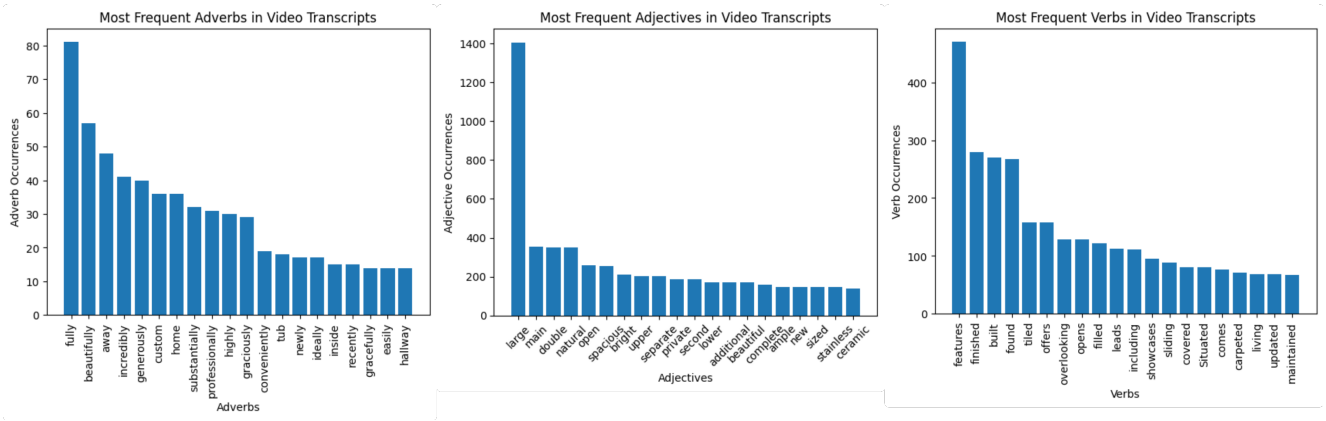



Table 4. **Constituency-based word frequency statistics.**

focus on quality materials and cozy elements. Decorative items like "chandeliers" highlight style and luxury in living spaces. Functional features such as "undermount sinks," "fridges," and "microwaves" underscore practical aspects of home living. The presence of "pot lights" and "California shutters" suggests attention to lighting and window treatments. Additional mentions of "stoves," "dishwashers," and "tile backsplashes" reflect both essential and aesthetic components of kitchens.

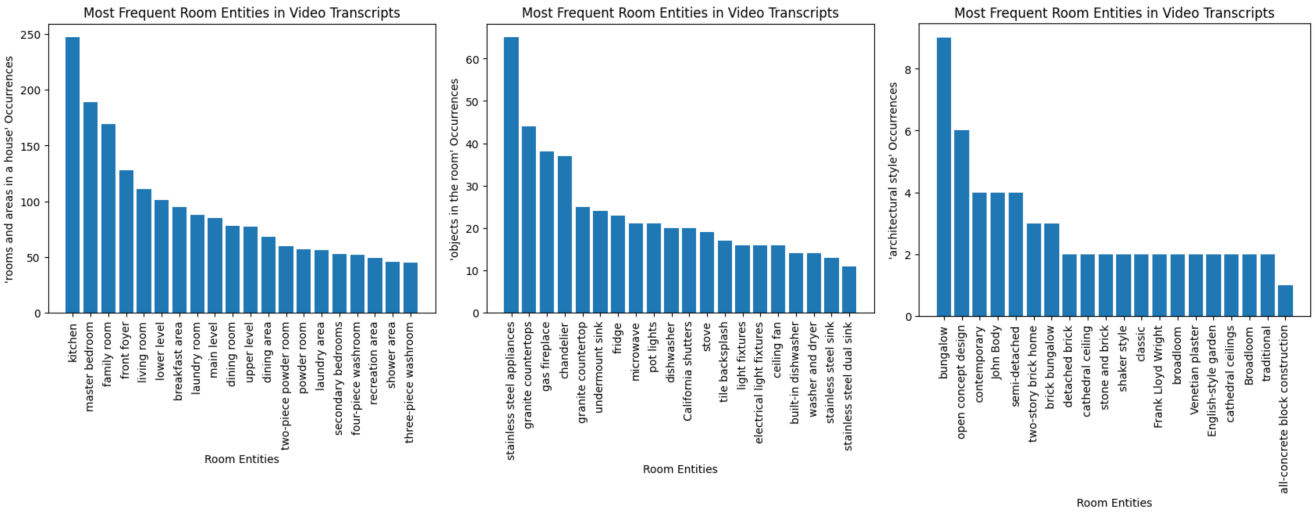



Table 5. Named Entity Frequency Statistics

The third and the last plot visualizes the *architectural style* found within the real estates. "Bungalow" emerges as the most frequently cited style, followed by "open concept design," indicating a preference for single-storey (usually plus the lower floor) living and spacious, flowing interiors. Styles like "contemporary" and "semi-detached" also receive notable mentions, highlighting a mix of modern and practical designs. The inclusion of "two-storey brick home" and "brick bungalow" suggests an appreciation for classic, durable architecture. Terms like "cathedral ceiling" and "stone and brick" emphasize distinct design features and materials. Styles attributed to iconic architecural figures like "Frank Lloyd Wright" reflect a nod to renowned architectural influences. The mentions of "English-style garden" and "Venetian plaster" suggest an interest in incorporating thematic and textural elements. Overall, this plot underscores a variety of styles, balancing traditional and modern influences in architectural preferences.

**Captures from the HouseTour Dataset.** In Figure 1, we provide some captures from the dense point clouds of our HouseTour dataset.

## 2. Additional Qualitative Results

Figure 2 provides additional qualitative results on trajectory generation from *Residual Diffuser*, and Figure 3 on scene-level summary generation from *Qwen2-VL-3D*.

## 3. User Evaluation of Text Generation

We conducted a single-blind user study in which three different participants evaluated generated descriptions for 20 different scenes. Across all assessed categories (Tab. 6), users consistently preferred our method. The results suggest that incorporating 3D positional information significantly enhances the perceived quality of the generated text. In our experimental setup, each participant watched a house tour video accompanied by two textual descriptions: one generated by a standard fine-tuned (SFT) baseline and the other by our Qwen2-VL-3D model. For every video, the order of the two descriptions was randomized to eliminate order bias. Six participants rated each description on a 0–10 scale, where 0 indicates no correspondence with the video and 10 indicates perfect alignment. The two descriptions were shown simultaneously, allowing for both absolute and comparative assessments. The definitions for each grading category is as follows:

| Methods | Lay. | Mat. | Fix. | Amb. | Ove. |
|---|---|---|---|---|---|
| Qwen2-VL-7B (SFT) | 6.1 | 6.0 | 5.9 | 5.9 | 6.0 |
| Qwen2-VL-3D (Ours) | **7.0** | **7.3** | **7.2** | **6.9** | **7.3** |

Table 6. **User Study** evaluating text generation quality across five categories: **Layout** (Lay.), **Material** (Mat.), **Fixture** (Fix.), **Ambience** (Amb.), and **Overall** (Ove.). Score range: [0,10].

**Layout** Does the description demonstrate a good understanding of the spatial organization of the room? Consider references to walls, doors, windows, room shapes, and how space is structured.

**Material** Does the description accurately capture the materials and finishes in the scene? Look for details about surface textures, colors, or types of materials (e.g., wood, glass, concrete).

**Fixture** How well does the description mention relevant built-in or fixed elements? Examples include lighting fixtures, sinks, cabinetry, or any permanent installations.

Figure 1. **Gallery of sample captures from the HouseTourataset**. The showcased regions originate from the 3D reconstructions.

**Ambience** Does the description convey the overall mood or atmosphere of the space? Consider lighting, color tone, and emotional or sensory impressions.
**Overall** An overall grading on the quality of the description.

## 4. Implementation Details

**Residual Diffuser.** We employ a lightweight U-Net architecture with two downsampling and two upsampling layers, changing the trajectory length by a factor of two at each

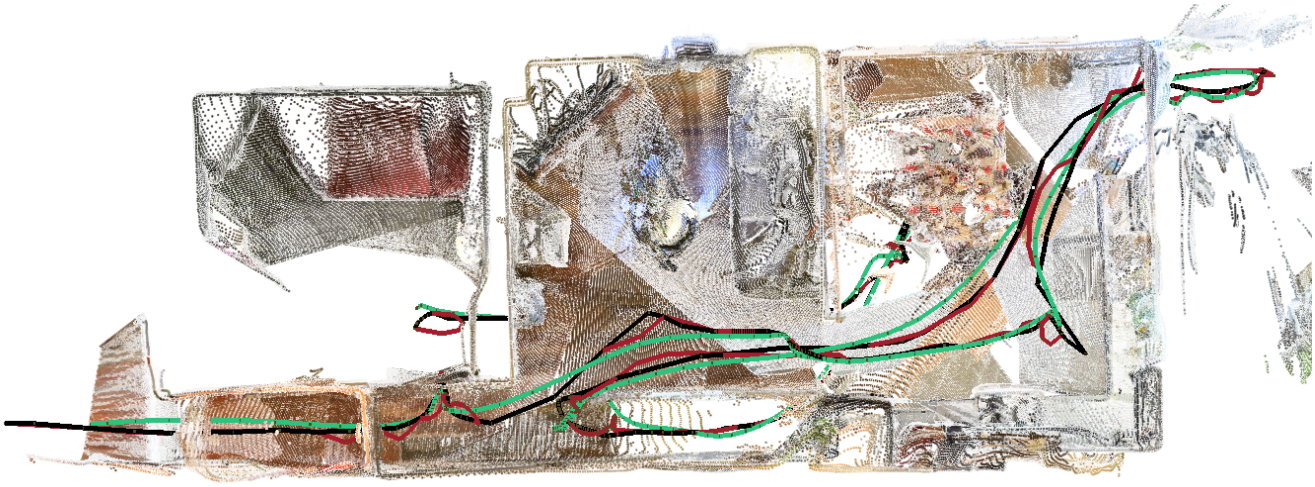


Figure 2. **Trajectory Visualization Within the 3D Reconstructions (top view).** Our method, Residual Diffuser, achieves a more human-like and smooth trajectory than the baseline Catmull-Rom Spline. **Black:** Ground-Truth, **Green:** *Residual Diffuser* and **Red:** Catmull-Rom Spline.

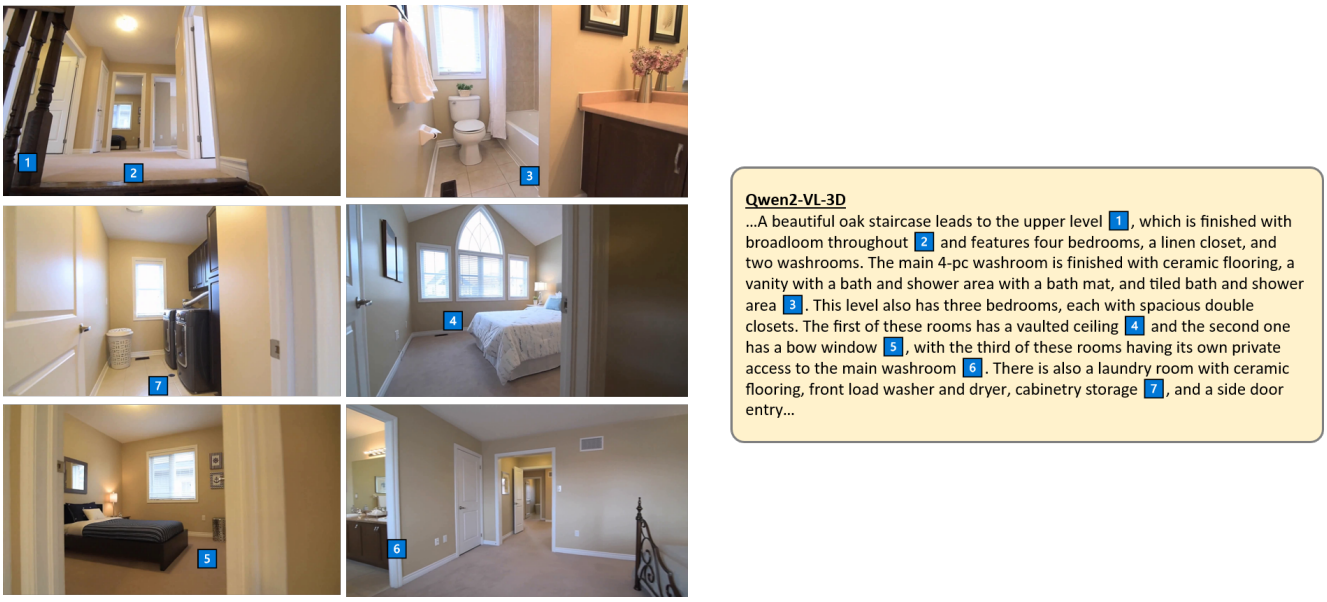



Figure 3. Further qualitative results for scene-level summary generation

step. The model is trained on a single *NVIDIA GeForce RTX 2080 (8GB)* GPU for 30K iterations with a batch size of 1 and gradient accumulation every 8th iteration, using a learning rate of $5 \times 10^{-6}$. To improve generalization, we randomly vary the number of sparse observations per training step, ensuring at least one observation every 20 frames.

**Qwen2-VL-3D.** We adopt a two-step training strategy for Qwen2-VL-3D. In the first phase, we LoRA-finetune Qwen2-VL on a single *NVIDIA A100 (80GB)* GPU for 20 epochs, with early stopping after the 8th epoch. We use the AdamW optimizer with a learning rate of $2 \times 10^{-4}$, cosine scheduling, and a warm-up ratio of 0.03. The effective batch size is 1, with gradient accumulation every 8th iteration. We train the VLM in *bfloat16* precision and apply gradient clipping at a maximum norm of 0.3 to prevent overflow. For the LoRA adapter, we specify both rank and alpha as 64, along with a 0.05 dropout rate. Adapters are added only to the Q and V weights of the attention layers.

**Inference Times.** All benchmarks were run on a 40 GB NVIDIA A100 GPU. The lightweight Residual Diffuser generates one trajectory in 0.23±0.04 s on average, whereas Qwen2-VL-3D requires 14.9±5.6 s to produce a single tex-

tual response.

## 5. Out-Of-Distribution Scenes

To evaluate the generalization capabilities of our method, we test it on two out-of-distribution (OOD) scenarios: (1) a single office room from the ScanNet++ dataset and (2) an online drone-view exterior shot of a property (Fig. 5. For these examples, we adjust the decoding temperature to $T = 0.3$ (compared to $T = 1.0$ used on the in-distribution HouseTour dataset). The temperature scaling sharpens the softmax distribution and, in practice, lets the weak image evidence outweigh generic language priors, dampening the language priors learned during training (e.g., persistent mention of bedrooms or kitchens in training samples); hence curtailing hallucinations that do not align with the visual evidence. While the tweak is effective, it also showcases the narrower visual-text diversity of the training data; expanding the dataset should further improve OOD robustness and lessen the reliance on temperature scaling. Note that, in both cases, the generated text is featuring the learned language style. As shown in Figure 4, the trajectory we generate (shown in blue) exhibits smoother and more natural motion compared to linear interpolation (red) and Catmull-Rom splines (green).

## 6. Evaluation In Metric Scale

In our experiments, we report all evaluation results in metric scale, even though our 3D reconstructions are not inherently metric. We adjust the scale of the reconstructions for evaluation so that errors in both large and small scenes are handled consistently, avoiding the distortion that arises from varying relative scales.

To achieve this, we use the same metric depth model employed by Mast3r [2] for metric alignment of camera pose pairs. Specifically, we uniformly sample 20 pairs of sequential keyframes from each reconstruction and measure the Euclidean distances between their poses as estimated by the Mast3r model. We then compare these distances with the corresponding Euclidean distances in our reconstructions. The ratio of the Mast3r-based distance to the reconstructed distance gives a scale multiplier for each pair, and we average these values at the scene level. The resulting mean scale multiplier for a scene is then applied to align that scene's trajectory to metric scale.

Figure 6 shows the mean and standard deviation of the scene-wise metric scale multipliers. As indicated by the figure, most of the scale multipliers fall within a reasonable range, with low variance.

## 7. Bradley-Terry Evaluation

In this section, we explain our evaluation process to generate Bradley-Terry (**BT**) normalized scores (between 0 and 1) for each of the Multi-Image-to-Text methods.

**Algorithm.** We begin by creating pairs of generated summaries, comparing outputs from each method to the ground truth for all 130 scenes in our test set. To mitigate positional bias, we shuffle these pairs to randomize the order in which they are presented to the LLM. For evaluation, we use the GPT-4o model [1] with a temperature of $0.5$ for text generation. Although we also tested open-source models like Llama-3.1-7b, we found they exhibit strong ordering bias, consistently favoring the first summary presented.

After gathering the binary preferences from the GPT-4o model, we construct a preference matrix $\mathcal{M}(i, j)$, where the entry in the $i^{\text{th}}$ row and $j^{\text{th}}$ column represents how many times method $i$ is preferred over method $j$. The score calculation algorithm starts with an initial guess for the parameters: each method is assigned a value of 1 (stored in the array $\pi_i$). These parameters represent the latent "strength" or ability of each method. We iteratively update the values of $\pi_i$ as follows:

$$\pi_i = \frac{wins_i}{\sum_{j \neq i} \frac{N[i,j]}{\pi_i + \pi_j}},$$

where $wins_i$ is the total number of wins for method $i$, and $N[i, j]$ is the total number of comparisons between methods $i$ and $j$. The denominator arises from setting the derivative of the likelihood function to zero (maximizing the likelihood), and it reflects how often method $i$ has been compared with method $j$ relative to their current estimated abilities. After updating all the parameters, the algorithm checks whether the maximum change in any parameter is below a specified tolerance threshold. If so, the iteration stops, indicating convergence.

**Transforming Parameters to Scores.** Once the parameters $\pi$ have been estimated, they are converted into scores $s_i$ in the interval $(0, 1)$ using the logistic transformation:

$$s_i = \frac{\pi_i}{\pi_i + 1}.$$

This transformation is useful because it maps the potentially unbounded $\pi_i$ values to probability-like scores, making them easier to interpret.

**Prompt to the Judge.** We use the following prompt as input to the GPT-4o judge:

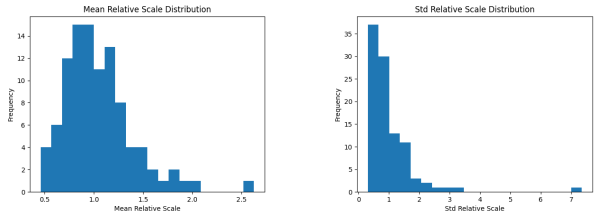
```
"Given Prediction 0, Prediction 1 and
the Ground-Truth texts, select which
text is closer to the ground-truth.
Evaluate the texts only based on the
information available in ground-truth."
```

"The spacious office features a large window with blinds for privacy and light control, as well as a whiteboard on the opposite wall for brainstorming sessions or team meetings. The open layout includes multiple workstations equipped with ergonomic chairs, computer monitors, and desktops. There is also a large desk area with additional chairs for collaborative workspaces. The office has ample storage options, including a wooden file cabinet, a credenza, and several deskside drawers. A large closet provides additional storage space, while a small closet near the door offers personal belongings storage. This room would be suitable for a variety of professional settings, such as law firms, accounting offices, architectural firms, engineering firms, and more."

Figure 4. **Inference on ScanNet++.** Generated text and trajectory based on 5 frames from scene: *56a0ec536c*.

"The home features a large front porch with a decorative ceiling and columns, as well as a front yard with a landscaped garden bed. The home has been freshly painted and features a new roof, new windows, and new flooring throughout. The home also includes a detached two-car garage with an attached carport, a large backyard with a garden bed, and a covered patio area."

Figure 5. **Inference on online video.** Generated text based on exterior drone shot, https://www.youtube.com/watch?v=QZrjZbI-H00.




(a) (b)

Figure 6. The **Mean** and **Standard Deviation** of the scale multipliers required to achieve metric scale.

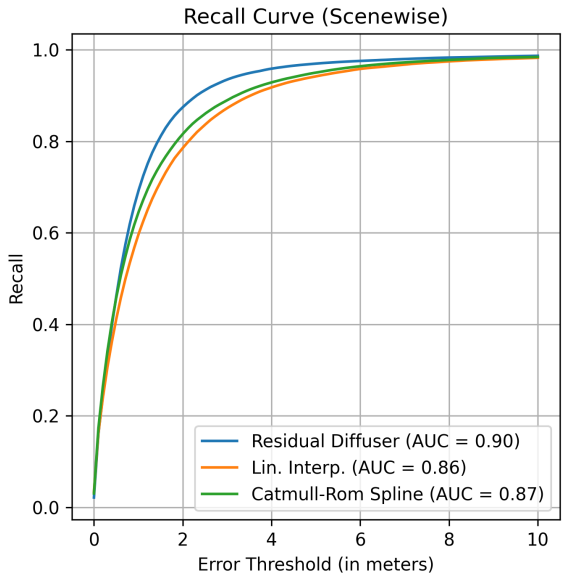



Figure 7. **Recall curve for 3D camera trajectory generation.** The x-axis shows the error threshold (in meters), and the y-axis indicates the ratio of predictions with errors below this threshold.

## 8. Recall Curves

In our trajectory generation evaluation, we use recall as a way of quantifying the magnitude of errors that the generation methods achieve. In the main paper tables (Tables 1 and 2) we report the results for *R@50cm*, *R@75cm* and *R@1m*. We provide the complete curves in Figure 7. Our method is shown to have the highest Area Under Curve (**AUC**) score and consistently shows better performance against the baselines with varying error thresholds.

## 9. Preliminaries

Generative modeling using denoising diffusion probabilistic models (DDPMs) aims to learn a probability distribution $p_\theta(\mathbf{x})$ that approximates the true data distribution of observed data $\mathbf{x}$. Unlike other generative methods – such as variational autoencoders or generative adversarial networks – which generate data in a single step, DDPMs gradually transform pure noise into structured data through an iter-

ative denoising process. The discrete stochastic denoising (reverse) process is modeled as a Markov chain, beginning at a predefined time step $T$ where the signal is considered to be pure noise $p(x_T) = \mathcal{N}(x_T; 0, I)$. A neural network $\epsilon_\theta$ is trained to predict the noise added at each timestep by minimizing the variational bound on the negative log likelihood, $\mathbb{E}[-log(p_\theta(x_0))]$. In practice, the reverse process is typically parametrized using Gaussian distribution as follows:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \quad (1)$$

Forward diffusion is a process that gradually adds noise to the data via a variance schedule $\beta_t \in (0, 1)$, determining the amount of noise introduced at each timestep $t$. This formulation enables a closed-form expression for sampling an arbitrary $x_t$, where $\alpha_t := 1 - \beta_t$ and $\bar{\alpha}_t := \prod_{0 \leq i \leq t} \alpha_i$. The conditional probability distribution $q(x_t|x_0)$ describes how likely $x_t$ is, given the clean signal $x_0$:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1 - \bar{\alpha}_t)I) \quad (2)$$

## References


[1] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 7

[2] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. pages 71–91, 2024. 1, 7

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[4] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR, 2023. 1

[5] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 1

[6] Johannes L Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I 13*, pages 321–337. Springer, 2017. 1